# The UTokyo System for Blizzard Challenge 2016

*Yi Zhao, Xiu You, Daisuke Saito, Nobuaki Minematsu*

Graduate School of Engineering, The University of Tokyo, Japan

{zhaoyi, youxiuenya, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

## Abstract

In this paper, we mainly introduce the UTokyo speech synthesis system for Blizzard Challenge 2016. Our system is a typical statistical parametric speech synthesis system. Its duration model is built by using HTS toolkit, and its acoustic model is made using Bidirectional Long Short-Term Memory with Recurrent Neural Network (BLSTM-RNN). In the synthesizing phase, sentence-level waveforms are firstly generated. Then these waveforms are concatenated into a paragraph. Because the evaluations of our system are not satisfactory, the defects and problems in the system are also discussed in this paper.

**Index Terms**: statistical parametric speech synthesis, HMM, BLSTM-RNN, audiobook

## 1. Introduction

The name of our team is 'UTokyo' and this is our first entry to Blizzard Challenge. The text-to-speech (TTS) system of UTokyo is built on hybrid models and belongs to the statistical parametric speech synthesis (SPSS).

Compared with the unit-selection method, SPSS is preferred because it can generate natural sounding synthetic speech with a rather small corpus [1]. Hidden Markov Model (HMMs) and Deep Neural Network (DNN) are two main approaches for acoustic modeling. HMM has been actively studied and various kinds of DNNs have significantly advanced the performance of SPSS recently [2, 3, 4, 5, 6, 7]. In our system, duration is modeled at HMM-state level, by a single Gaussian distribution for each state [8]. A hybrid structure of DNNs proposed in [4] is employed as acoustic model to predict the acoustic features of Mel-Generalized Cepstral (MGC), fundamental frequency (F0), band aperiodicity parameters (BAP) and unvoiced/voiced decision(UV).

Besides speech waveform generation, text analysis is another important component of a typical TTS system. British English is the given language of this year's challenge. The text analysis for British English was the most difficult problem we faced due to our lack of knowledge in it. Finally we had to choose an open source software.

Since our system didn't achieve satisfactory evaluations, we will try to analyze the defects or problems in it. We hope that the problems we faced and solutions we found may provide some useful information for other studies.

In this paper, a brief explanation to the task of this year's Blizzard Challenge is given firstly. The description of our system as well as techniques we adopted is shown in section 3. Subjective listening test results are displayed in section 4. Moreover, the defects in our system are analyzed and discussed in section 5. And conclusion and future work are presented in section 6.

## 2. Task Description

The task of Blizzard Challenge in this year is to produce a set of voices given some British English corpora. This database has approximately 5 hours' speech data. These speech data are recorded by one and the same female speaker. The sampling rate is 44.1kHz. In this database, there are three types of audio formats, that are mp3, wma and m4a.

The speech data are from 50 UK AudioBooks specially designed for English children from elementary to advanced levels. Among all the stories, 40 books are segmented chapter by chapter, and a sentence-level alignment label between text and speech is provided. For the other 10 books, only audios of the whole stories and picture books in PDF format are provided.

Many utterances in those audio books are very rich in emotion with a large number of onomatopoetic words. They also include non-speech sounds such as ringing and animal crying. In some cases, these sounds are overlapped with spoken utterances. The testing transcriptions given by the organizer include texts collected from audiobooks, news and semantically unpredictable sentences (SUS). Four different sets of audio are required to be synthesized. These sets are books, chapters, pages and lines.

## 3. System Description

The training framework of our system is shown in Fig. 1 and synthesis module is shown in Fig. 2. At the training stage, both texts and speech data are firstly pre-processed. After pre-processing, context-dependent linguistic labels are estimated from texts, and acoustic features are extracted from speech. By taking use of linguistic labels and acoustic features, single Gaussian distribution based duration distributions of HMM states are estimated, and Bidirectional Long Short-Term Memory with Recurrent Neural Network (BLSTM-RNN) based acoustic model are trained. At the synthesis stage, the testing transcriptions are pre-processed and context-dependent label sequences are estimated. Then duration parameters are determined based on the state-duration distributions and acoustic parameters are produced by the well-trained neural network. And next post-processing is done for predicted acoustic parameters. Finally waveforms are synthesized through a vocoder.

### 3.1. Data Preparation

The published data of Blizzard Challenge are quite different from what we generally used for speech synthesis studies. The speaking style of most utterances are neutral but some are very expressive. For some utterances that are transcribed with double quotation marks,the speaker always tries to generate unique characters in her voices including different speaker identities, different speaking styles, and so on. There are also some non-speech sounds in the recording, such as screaming, crying, ring-

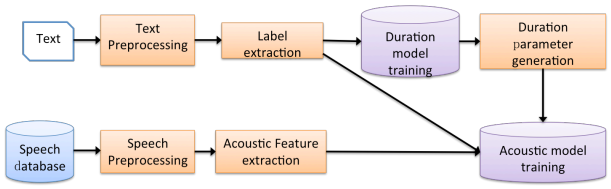


Figure 1: *Training of UTokyo speech synthesis system*

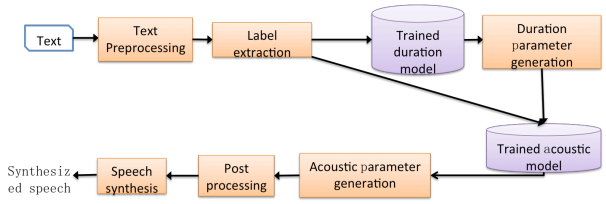


Figure 2: *Synthesis of UTokyo speech synthesis system*

ing and animal yelling. Transcriptions are not always consistent with utterances. And there are ten books which do not have corresponding text files. The audios are recorded in three different file formats. Their volumes and gains are different from each other. For all the texts and speech files, they had to be processed and aligned before building a system.

The procedure of our preprocessing heavily relies on human resources and it is quit laborious. Firstly, the transcriptions of ten stories are extracted from original picture books. Then, audio files and transcriptions of all the fifty stories are segmented and aligned sentence by sentence. In the meantime, non-speech sounds are removed, and wrong transcriptions are corrected. These tasks can be done automatically for the forty books because time labels are provided, but for the other ten books, these had to be done manually. Note that we kept the onomatopoetic words although their transcriptions are not accurate. Finally, all the speech data are normalized into 'wav' format. Their gains and volumes are normalized.

### 3.2. Linguistic & Acoustic Feature Extraction

Festival [9] is adopted for text analysis in our system. We use its default pronunciation dictionary, the Carnegie Mellon University Pronunciation Dictionary (CMUDICT) [10], for lexical look-up as well as stress assignment. Although CMUDICT is designed for North American English, we still use it because we don't have a better choice. Both monophone alignment labels and full context labels are extracted from the "utts" files generated by Festival.

Acoustic features including 39-dimensional mel-cepstral coefficients, F0 in log-scale, 26-dimensional Band-aperiodicity parameters (BAP), and their delta and delta-delta features are extracted with the help of STRAIGHT [21]. A binary value for voiced/unvoiced decision is also estimated.

### 3.3. Acoustic Model

The acoustic model we built are based on a hybrid deep learning structure which is mentioned in [4]. The neural network of our system has five layers, 3 lower layers are feedforward layers and 2 higher layers are BLSTM-RNN layers. Each layer has 300 nodes. Linguistic features are converted to 305 dimensional vectors. State index and frame index are also attached to the end of the vectors. Before training the neural network, state-level alignment is done by using HTS. Linear interpolation of F0 is applied over unvoiced segments. To train the acoustic model, both input and target features are normalized to zero mean and unit variance. Implementation of the network training is done with the help of a machine learning library "CURRENNT" [11].

### 3.4. Duration Model

Duration model is built with the help of HTS 2.3 [12]. It uses a semi-Markov structure in which the temporal structure is approximated by a Gaussian distribution [13]. To build the model, a five state left-to-right context dependent HMM is adopted. States of the context dependent HMMs are clustered by using a decision-tree based context clustering technique, and the tied context dependent HMMs are reestimated with the embedded training [14]. State durations are modeled by Gaussian distributions. Each HMM also has its explicit state-duration probability distribution to model the temporal structure of speech.

### 3.5. Speech Synthesis

Before synthesis, given paragraph-based transcripts are segmented into sentences. To synthesize speech, each of the sentence-based transcripts is normalized and converted to a context-based label sequence by using Festival in the first phase. In the second phase, a sentence HMM is constructed by concatenating context dependent HMMs according to the label sequence. State durations of the sentence HMM are determined based on the state-duration densities. Then, the label sequence is converted to a numerical vector, to the end of which, the predicted duration information is attached. By using the vector as input to the well-trained acoustic models, the required acoustic features are produced. Before waveform generation, global variances are combined with Maximum Likelihood Parameter Generation (MLPG) algorithm to enhance the dynamic properties of synthetic speech. To calculate the global variances, the variance of each sentence's acoustic features is built as a single GMM. At last, speech waveform is generated using the STRAIGHT vocoder. For paragraphs, pages and the whole story synthesis, we simply concatenate the synthesized sentences together.

## 4. Results and Analysis

In this section, we will discuss the evaluation results in detail. Our designated system identification letter is 'N'. System A is natural speech. System B is the Festival benchmark system based on unit-selection. System C is the HTS benchmark. System D is a DNN benchmark and others are participants systems. The subjects who are involved in the listening test are paid listeners, speech experts, and online volunteers.

This year, there are mainly four sections to evaluate, , which are a paragraph test, a naturalness test, a similarity test, and a SUS test. For the paragraph test, there are seven kinds of tests to evaluate different aspects of synthesized paragraphs, namely overall impression, pleasantness, speech pauses, stress, intonation, emotion, and listening effort.

For the paragraph test, our system seems to be one of the worst systems in terms of listening effort, pleasantness and overall impressions. Although the MOS score for speech pauses, intonation, stress, and emotion are not the lowest, the

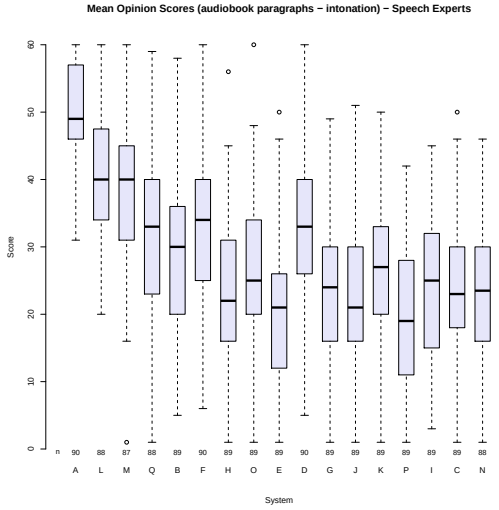


Figure 3: *MOS results on intonation of audiobook paragraphs*

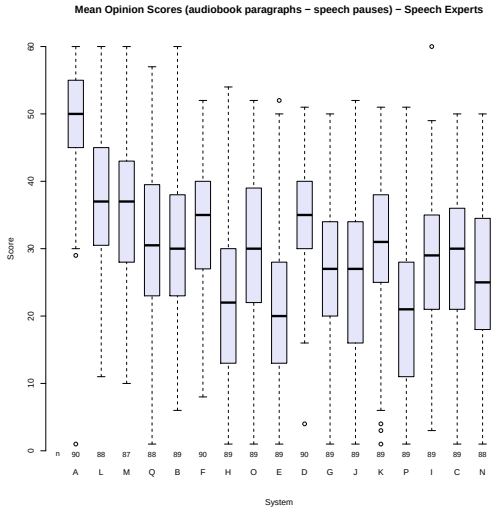


Figure 4: *MOS results on speech pauses of audiobook paragraphs*

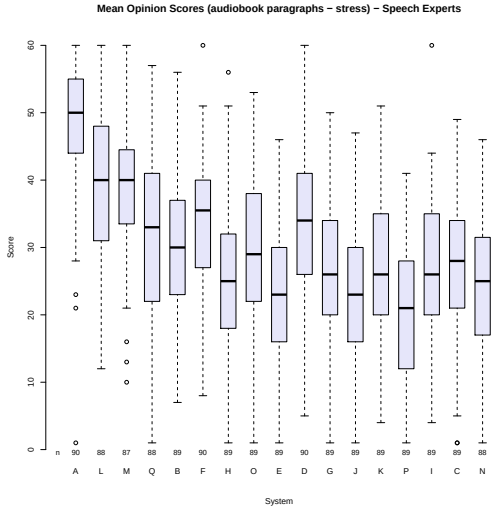


Figure 5: *MOS results on stress of audiobook paragraphs*

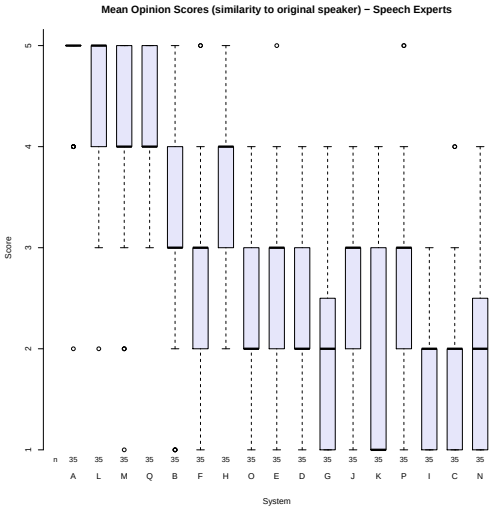


Figure 6: *MOS results on similarity*

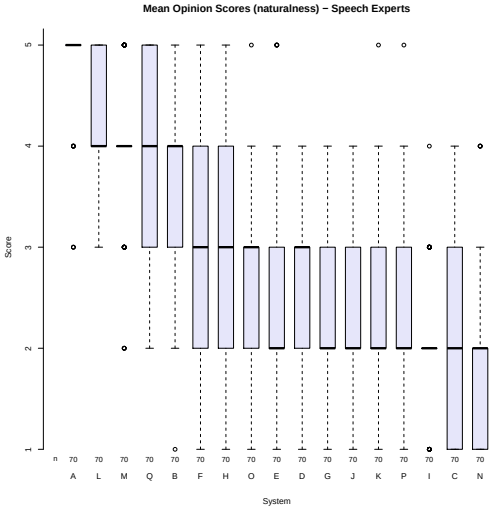


Figure 7: *MOS results on naturalness*

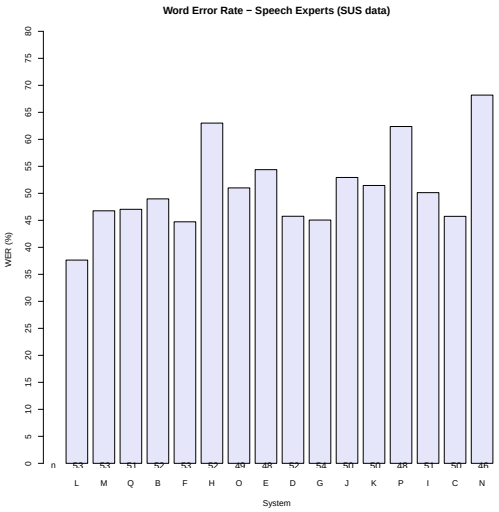


Figure 8: *Results of word error rate of SUS data*

evaluations are still poor. Our naturalness result, similarity result and word error rate of SUS are also very poor. We list out some evaluation results of speech experts in Fig. 3 to Fig. 8. In the next section, we will mainly analyze the defects in our system.

# 5. Defects in UTokyo System

In this section, we mainly discuss the problems in our system.

## 5.1. Defects in Data Processing

1) Some audios and transcriptions are aligned and segmented manually, which is laborious and time consuming, and sometime errors might happen. We hope to build a speech recognition system next time.

2) Dictionary selection. We use "cmudict" for lexical look-up as well as stress assignment. However, "cmudict" is a North American English dictionary while the training data are British. In the meantime, although the pronunciation of provided data is British English, the transcriptions are written in American English. We consider this will degrade the quality of synthesized speech to some degree. We also tried 'OALD' dictionary, but the effect is quite weak.

3) Data selection. Some words do not have accurate transcriptions, such as onomatopoetic words and thrilling. We kept them in our training data. But this is inappropriate.

4) Context labels. The speaker imitated several different speakers' voice, but we didn't take their differences into consideration. In future work, rich labels are examined for expressive synthesis.

5) Mono-phone alignment. Alignment results of mono-phone labels are not totally accurate by using Festival.

## 5.2. Defects in Model Training

1) Neural network selection. We only tried several network structures. In fact, we are not sure whether we have found the optimum neural network structure for acoustic parameter generation or not.

2) Duration model. The duration model is trained using all of the training data. However, the speaking rate and intonation vary from sentence to sentence. It's better to gather utterances into several clusters according to the speaking rate.

## 5.3. Defects in Speech Generation

1) Paragraph-level speech synthesis. In our experiments, paragraph-level speeches are generated by concatenating sentence-level speeches, which ignored the importance of pauses between sentences.

# 6. Conclusion

We introduced the UTokyo speech synthesis system for Blizzard Challenge 2016. The results of listening test for our system are not good, but we have found many interesting problems that we should have attacked. We will try to build a better system for next years Blizzard Challenge.

# 7. References


[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *ICASSP*. IEEE, 2013, pp. 7962–7966.

[3] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *ICASSP*. IEEE, 2015, pp. 4470–4474.

[4] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks," in *Interspeech*, 2014, pp. 1964–1968.

[5] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (dnn) for parametric tts synthesis," in *ICASSP*. IEEE, 2014, pp. 3829–3833.

[6] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis," in *ICASSP*. IEEE, 2015, pp. 4475–4479.

[7] Y. Q. Yuchen Fan, F. K. Soong, and L. He, "Unsupervised speaker adaptation for dnn-based tts synthesis," in *ICASSP*. IEEE, 2016, pp. 5135–5139.

[8] Z. Heiga, K. Tokuda, T. Masuko, T. Kobayasih, and T. Kitamura, "A hidden semi-markov model-based speech synthesis system," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 825–834, 2007.

[9] P. Taylor, A. W. Black, and R. Caley, "The architecture of the festival speech synthesis system," 1998.

[10] R. Weide, "The carnegie mellon pronouncing dictionary [cmudict. 0.6]," 2005.

[11] F. Eyben, J. Bergmann, and F. Weninger, "Currennt cuda-enabled machine learning library for recurrent neural networks."

[12] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[13] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-markov model based speech synthesis." in *INTERSPEECH*, 2004.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for hmm-based speech synthesis." in *ICSLP*, vol. 98, 1998, pp. 29–31.